Syst. Biol. 58(2):167–183, 2009

DOI:10.1093/sysbio/syp018
Advance Access publication on June 12, 2009

A Method for Investigating Relative Timing Information on Phylogenetic Trees

DANIEL FORD[1], FREDERICK A. MATSEN[2*], AND TANJA STADLER[3]

[1]*Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA;*
[2]*Department of Statistics, University of California, Berkeley, 367 Evans Hall #429, Berkeley, CA 94720-3860, USA;*
[3]*Institut für Integrative Biologie, ETH Zentrum/CHN K 12.2, Universitätstrasse 16, 8092 Zürich, Switzerland;*
**Correspondence to be sent to: Department of Statistics, University of California, Berkeley, 367 Evans Hall #429, Berkeley, CA 94720-3860, USA; E-mail: matsen@berkeley.edu.*

All authors contributed equally to this manuscript.

Abstract.—In this paper, we present a new way to describe the timing of branching events in phylogenetic trees. Our description is in terms of the relative timing of diversification events between sister clades; as such it is complementary to existing methods using lineages-through-time plots which consider diversification in aggregate. The method can be applied to look for evidence of diversification happening in lineage-specific "bursts", or the opposite, where diversification between 2 clades happens in an unusually regular fashion. In order to be able to distinguish interesting events from stochasticity, we discuss 2 classes of neutral models on trees with relative timing information and develop a statistical framework for testing these models. These model classes include both the coalescent with ancestral population size variation and global rate speciation–extinction models. We end the paper with 2 example applications: first, we show that the evolution of the hepatitis C virus deviates from the coalescent with arbitrary population size. Second, we analyze a large tree of ants, demonstrating that a period of elevated diversification rates does not appear to have occurred in a bursting manner. [Branch length; key innovation; neutral models; phylogenetics.]

Understanding the tempo and mode of diversification is one of the major challenges of evolutionary biology. Phylogenetic trees with timing information are powerful tools for answering questions about tempo and mode. Such trees were once available only in situations with a rich fossil record, where the timing information might have come from radiocarbon dating or stratigraphic information. However, modern techniques of phylogenetic analysis not only are capable of reconstructing the topology of phylogenetic trees, but also can reconstruct information about the timing of diversification events even when limited or no fossil evidence is available. This can be done in one of a number of ways. One can first test if a molecular clock is appropriate (see Felsenstein 1981, 1988) and then reconstruct under the assumption of a molecular clock. Or one can reconstruct a tree with branch lengths using any method and then apply rate smoothing (Sanderson 2003). One may also choose from the variety of "relaxed clock" methods that allow the rate of substitution to vary within the tree (Gillespie 1984; Huelsenbeck et al. 2000; Drummond et al. 2006). Of course, the accuracy of any of these techniques depends on a correct choice of model and a strong phylogenetic signal, along with perhaps some fossil calibration points.

Phylogenetic trees with timing information can then be used to make inferences about the forces guiding the evolution of the taxa. For example, the paper by Moreau et al. (2006) observes that there was a period of high diversification rate in ant lineages during the rise of angiosperms. Another paper by Harmon et al. (2003) uses the deviation of 4 groups of lizards from the pure-birth (i.e., constant rate of speciation across lineages; no extinction) model of diversification to make inferences about their evolutionary radiations.

One way to compare trees to models is based on likelihood and model selection. The strategy of these tests is to calculate the likelihood of several models of varying complexity and then to choose the model with the best balance of likelihood and complexity. Paradis (1997) initiated this approach, choosing between 3 diversification models based on a likelihood ratio test and the Akaike Information Criterion. More recently, Rabosky (2006) and Rabosky and Lovette (2008a, 2008b) have furthered this work, choosing between various models of global rate variation.

Our work presented here is different from that of these authors in several respects. First, although we will call rejection of neutral models "lineage-specific bursting" (LSB) speciation, the method is not tied to any specific alternative diversification model. Rather, we are proposing the use of a summary statistic, analogous to γ (described below), or a tree shape statistic, which can be shown to have a known distribution under certain types of models and thus can be used to test deviation from those models. If a new type of neutral model is developed, our statistics may well prove useful to test that model.

Because our method does not compute fit in terms of likelihood, it can be used to compare the fit of models with an arbitrary number of fit parameters. In contrast, because one must discount the likelihood by a function of the number of free parameters when comparing the fit of various models using likelihood, the models compared using likelihood tend to have a small number of parameters. Paradis (1998), for instance, allows for shifts in diversification parameters along several edges of a phylogenetic tree. In the alternative model of Rabosky (2006), there is a change in the speciation and extinction rates for all lineages at a single time.

Rabosky and Lovette (2008b) consider demographic scenarios where speciation and extinction rates follow constant or exponential trajectories. Our methods, on the other hand, can imply rejection of whole classes of models with an arbitrary number of parameters, such as the complete class of models with global speciation and extinction rates.

A simpler approach is taken by Nee, May, and Harvey (1994), who test the hypothesis of equal speciation and extinction rates for a collection of lineages in the tree. The test works as follows: assume that there are some number n of descendants of k lineages. Under the hypothesis of equal rates, one can obtain closed-form equations for the probability that one of these lineages had q descendants or more. Given some tree, one can then attach a p value to the observed number of descendant lineages of the k chosen lineages under the equal rates assumption. For more details, see Schluter (2000).

Another popular way of comparing trees to models uses lineages-through-time (LTT) plots and the associated γ-statistic introduced by Pybus and Harvey (2000) (for a helpful review article, see Ricklefs 2007). LTT plots have time t on the x-axis and simply show the number of lineages that were present in the phylogenetic tree at time t on the y-axis. A constant rate pure-birth process would have the number of taxa increasing exponentially; it is thus common to compare LTT plots to an exponential curve (Zink and Slowinski 1995; Harmon et al. 2003). The γ-statistic is computed from the length of the periods during which the number of lineages stays constant (called the "internode intervals"): if g_i is the time between the $i-1$th branching event and the ith branching event for $i = 2, \ldots, n$, then

$$\gamma = \frac{\sum_{i=2}^{n-1} \sum_{j=2}^{i} j g_j - \frac{t}{2}}{(n-2)\frac{t}{\sqrt{12(n-2)}}},$$

where t is the tree length $\sum_{j=2}^{n} j g_j$. The γ for a pure-birth diversification process will have a standard normal distribution. Broadly speaking, $\gamma < 0$ implies that diversification rates were high early in history, whereas under the pure-birth process $\gamma > 0$ implies that most diversification has happened more recently. A similar statistic with the same goals in mind was constructed by Zink and Slowinski (1995).

However, much more information is available in a phylogenetic tree with diversification timing information than that which can be summarized in an LTT plot or a derivative statistic. Consider the tree in Figure 1, with 2 sets of sister taxa, A and B. The taxa in B had a period of relatively high diversification rate early in evolutionary history, during which time the lineage leading to A is in a period of stasis. Then lineage A experiences a burst of diversification, and the taxa in B do not experience any bifurcation (lineage-splitting) events during this time. We will call the sort of diversification seen in Figure 1 LSB diversification, as such patterns can emerge when lineages descending from a single node alternate periods of high diversification. Note also that we will often use the word "bifurcation" of lineages rather than "speciation", as the former word pertains to both the micro- and the macroevolutionary settings.

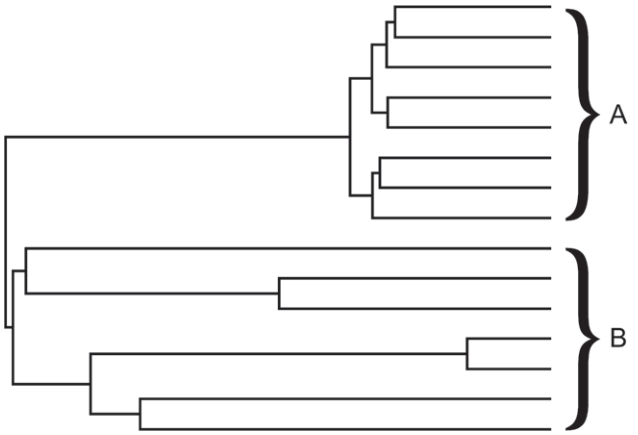

FIGURE 1. A motivating example showing "bursting" diversification. Namely, in the oldest part of the tree, diversification events happen exclusively in the B lineage, followed by a period of high diversification rate in the A lineage. This paper constructs a statistical framework for analyzing such "bursting" patterns or their opposite.

The LSB diversification seen in Figure 1 would not be apparent in an LTT plot. Indeed, LTT plots take the timing information out of the context of the phylogenetic tree from which they are derived and thus ignore information about how the timings relate to topology of the tree. This context can be crucial, as we now argue.

One would like to be able to say if, for example, the pattern seen in Figure 1 arose simply "by chance." In order to do so, we need 2 things: first, a convenient way to summarize the timing information and, second, a set of neutral models which define what we mean with "by chance." For a given internal node, we summarize the relative timing information at that node by writing down the order of diversification events by clade. For instance, we associate with the root node of Figure 1 the sequence $s = BBBBBAAAAAAAB$ which we will call a "shuffle" in analogy to a shuffling of cards labeled A and B. We make a more formal definition of shuffles in the section labeled "Tree Shuffles."

Now that we have summarized the relative timing information as a shuffle, s, at the root node we would like to think about if s arose "by chance." This of course requires us to define a probability distribution on shuffles; we demonstrate below that a wide class of neutral models on phylogenetic trees gives the uniform distribution on shuffles. The uniform distribution in this setting is what one would get by throwing the A's and B's of the shuffle into a bag and drawing them out one by one uniformly at random. Thus, it seems reasonably unlikely that the shuffle s would arise by chance, having first a long run of B's and then a long run of A's.

We can attach a p value to a shuffle by using the "runs distribution." The number of "runs" is simply the number of sequences of the same letter: in this case, there is a run of B's, then a run of A's, then another of B's.

That totals 3 runs. Under the uniform distribution, the probability of seeing a given number of runs in this setting is known from classical statistics and can be calculated via equation (1). The probability of seeing 3 runs with 6 *A*'s and 7 *B*'s is about 0.00641, and the probability of seeing 2 runs is about 0.00117. We can interpret the sum of these 2 probabilities, 0.00758, as the significance level of the LSB diversification seen in Figure 1. Being below the 1% significance level, we can interpret this shuffle as being quite significant; thus, if the tree in Figure 1 came from data, the observed lineage-specific diversification might require some explanation.

One can use shuffle-derived information in 2 ways. First, one can look at a single node and consider the significance level of that node as done in the above example. Such an approach would make sense if one has a prior hypothesis about a single clade, for example, if one would like to test if a given innovation led to an increased level of diversification in the short term. However, care needs to be taken to avoid a multiple testing problem; for example, a statistically dangerous approach would be to use the shuffle p values of all the internal nodes to find one which looks like it is "bursting" and then use that p value without a correction.

Alternatively, we investigate all shuffles simultaneously in the rest of this paper by taking the sum of the number of runs for all internal nodes in the tree. In doing so, a statistically significant total number of runs may indicate that diversification is driven by "repeated" discovery of new key innovations. Viruses provide an example of such a process: when a virus mutates in a significant way to escape the host's immune defenses, it gains an advantage over the rest of the viral population. This may allow it to increase to a higher frequency, and in doing so it diversifies in other ways; upon reconstruction, such diversification should be visible as a collection of lineage-specific bursts. In macroevolution, one would expect to see such a pattern when key innovations repeatedly open up new ecological niches.

However, like a significantly negative value of γ, a significantly small value of the runs distribution may appear for a number of reasons. Certainly a substantial change in relative diversification rates would lead to clustering of the same letter, but the converse need not hold. For example, assume in the macroevolutionary case that there are 2 lineages descending from the root; the left lineage has a moderate speciation rate and no extinction, whereas the right lineage has a very high speciation and extinction rate. Because of the high extinction rates, the right lineage will have most of its internal nodes at a time close to the present day. This phenomenon has been called "the pull of the present" (Nee, Holmes, et al., 1994). Such a setting might be identified by our method as LSB diversification, even though no change in relative diversification rates has occurred.

Is the "pull of the present" likely to cause a false identification of varying relative diversification rates? It is possible, but consider what would be involved. Because in our applications we are taking the sum of the number of runs across all internal nodes of the tree, it is not enough to have just a single pair of sister clades with significantly differing extinction rates: there must be a general pattern of such extinction rate differences across the tree. Thus, consider a phylogeny with 2 subtrees, *L* and *R*, branching off the root. Say subtree *L* has 2 subtrees *LL* and *LR*. In order to have a nonuniform shuffle due to the pull of the present, we need to have the extinction rate in (say) *L* be significantly higher than that in *R*, and within *L* we need the extinction rate of (say) *LL* to be significantly higher than that of *LR*. Continuing in this way down the tree, there must be some lineages with exceedingly high bifurcation and extinction rates compared with others. Although such a setup is possible, we consider it to be less likely than changes in relative diversification rates.

We also note that this method, like any method based on phylogenetic trees, is subject to biases introduced by nonuniform taxon sampling. As discussed below, the runs p value is not biased by uniform taxon sampling; however, it is not hard to devise a sampling scheme which would bias the results. For example, say we have 2 subpopulations descending from a single internal node by a process that induces the uniform distribution on shuffles. On side *L* sampling is done uniformly, whereas on side *R*, similar lineages are unlikely to be part of the sampling. Such a scheme would bias the surviving internal nodes in *R* to be farther back in the past, resulting in a nonuniform distribution on shuffles.

We now present the goals and scope of this paper. The first aim of this paper is to provide analytical tools to compare patterns of diversification between lineages. In doing so, we hope to provide a complementary perspective to that provided by LTT plots and associated statistics. In particular, we would like to detect cases where the relative diversification rates in 2 sister clades vary over time. One might expect changes in diversification rates if a lineage diversifies to fill variants of a single niche or if a key innovation appears that makes further diversifications more likely. By comparing the results of our analysis to results using LTT plots, we may be able to tease apart causes of diversification rate changes—are they lineage specific or due to global events?

The second aim of this paper is to develop neutral models of phylogenetic trees with relative timing information. In contrast to the setting of phylogenetic tree shape, where a number of models are available (Aldous 1995; Ford 2006; Mooers et al. 2007), there are relatively few models available for trees with any notion of timing information. Null models are important as they allow us to distinguish between stochastic sampling and actual events which need investigation; they are thus important tools for assessing significance.

We conclude the paper with example applications. Our first example application uses hepatitis C virus (HCV) data and shows that trees from these data demonstrate a limited but significant amount of LSB diversification. Furthermore, several of the HCV data sets show a very substantial deviation from the coalescent model as seen by the runs statistic and a classical tree shape

statistic. This analysis implies a note of caution for researchers using coalescent methods to analyze HCV data. Our second application is to the ant data of Moreau et al. (2006) and Moreau (2008), the lineages of which do not appear to demonstrate significant LSB diversification, despite some other interesting characteristics of their history.

Our paper is one contribution to the area of understanding mechanisms of diversification from phylogenetic trees, which is a very large field of study. Besides LTT plots, γ, and likelihood methods, there is an entire literature on phylogenetic tree shape (which does not include branch length); for an excellent review, see Mooers and Heard (1997). There are also a number of interesting papers which use trait or geographic information, for example, Pagel (1997), Ree (2005), Weir (2006), and Maddison et al. (2007).

Tree Shuffles

Our method is based on "ranked" phylogenetic trees: trees for which the order of branching events in the tree is specified in a way compatible with the topology (more specific definition below). Such trees have been called "dendrograms" (Page 1991). We will show that a ranked phylogenetic tree is equivalent to a phylogenetic tree with a "shuffle" at each internal node specifying relative timing information. As described below, a broad class of neutral diversification models give the uniform distribution on shuffles, which leads to some natural tests for deviation from these models. Thus, evidence of deviation from the uniform distribution on shuffles is evidence of deviation from this entire class of neutral models. (Often, a model with branch lengths is given, in which case we consider the induced model given by considering ranks.)

The intuition behind the shuffle idea is presented in Figure 2. The relative order of bifurcation events for an internal node of a tree is determined by the sequence of full and hollow circles on the left side of each tree. We call this sequence a "shuffle." Shuffles also have a natural interpretation in terms of evolutionary history. Namely, "bursting" diversification leads to symbols of a shuffle clustering together. The opposite situation, where there is a postdiversification delay before a lineage can diversify again, can be recognized by the interspersing of different symbols. This latter situation has been called "refractory" diversification (Losos and Adler 1995).

We now make more formal definitions of our terms. For the purposes of this paper, a "phylogenetic tree" is a rooted tree with distinct leaf labels. We assume that the trees are equipped with branch lengths which make them ultrametric (i.e., the total branch length from the root to any leaf is constant). We will also assume that the tree is only on extant taxa, that is, that no extinct taxa are included in the tree. We will denote the set of interior nodes of a phylogenetic tree T with N_T. For an internal node v in N_T, define T_v to be the rooted subtree of T containing all the descendants of v. The "daughter trees" of v are the 2 subtrees of T_v which we obtain by deleting v and its 2 incident edges. For the first part of the paper, we assume that our phylogenetic trees are bifurcating (and describe later how to generalize the ideas presented to the case of multifurcating trees).

A "rank function" on an arbitrary set S is simply an ordering of the elements of that set; mathematically, it is a one-to-one mapping from S to ranks $\{1, 2, \ldots, |S|\}$. A "rank function on a phylogenetic tree" T is a rank function on the set of interior vertices N_T with the property that the ranks are increasing on any path from the root to a leaf. We call a phylogenetic tree with a rank function a "ranked phylogenetic tree" or simply a "ranked tree" (Semple and Steel 2003). Assuming that no 2 bifurcation events happen simultaneously, any clock-like phylogenetic tree will define a unique ranked tree; the order in the ranking is given by time.

In this paper, an (m, n) *shuffle* on symbols p and q is simply a sequence of length $m + n$ containing m p's and n q's. The complete terminology for such a sequence is "riffle shuffle" (Aldous and Diaconis 1986). For example, $pqppq$ is a $(3, 2)$ shuffle on p and q.

We can use shuffles to develop a recursive formulation of ranked phylogenetic trees. Assume that v is an internal node of a tree and that the tree T_v containing the descendants of v is composed of 2 daughter subtrees L_v and R_v. Assume that L_v and R_v have m and n internal nodes, respectively. We define a "shuffle at an internal node" v to be an (m, n) shuffle on symbols ℓ and r. The utility of these shuffles in the present context is summarized in the following observation.

Observation 1. *Let T be a tree with the 2 distinguishable daughter trees L and R. Let L (respectively R) have m (respectively n) interior vertices. Given a rank function on L and R,*

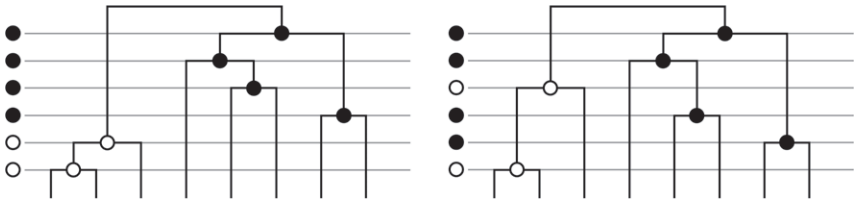

FIGURE 2. A shuffle at a given internal node. Bifurcations on the left subtree are marked with a hollow circle, and those on the right subtree are marked with a solid circle. The relative timing for these events is shown beside the tree; we call this sequence of symbols a "shuffle." A set of shuffles for every internal node of a phylogenetic tree exactly determines the relative order of bifurcation events. Similar type symbols occurring together as in the left tree is evidence of lineage-specific bursts.

the rank functions on T respecting the rank functions on L and R are in one-to-one correspondence with the (m, n) shuffles on the symbols l, r.

To see how this works, assume that orderings $l_1 < l_2 < \cdots < l_m$ on the interior vertices of L and $r_1 < r_2 < \cdots < r_n$ on the interior vertices of R are given, along with an (m, n) shuffle on l and r. The required ranking on T is obtained by progressing along the shuffle and substituting l_i and r_j for l and r in order: for example, the shuffle *lrllr* uniquely defines the ordering $l_1 < r_1 < l_2 < l_3 < r_2$ when $l_1 < l_2 < l_3$ and $r_1 < r_2$. In the other direction, a rank function on T uniquely defines an (m, n) shuffle and a rank function on L and R.

By this observation, a rank function on the internal nodes of T_v respecting the rank functions on the internal nodes of L_v and R_v is equivalent to a shuffle at the internal node v. Therefore, we can recursively reconstruct the rank function for any ranked oriented tree (ROT) given a shuffle at each internal node. We define a *tree shuffle* to be such a choice of shuffles. With Observation 1, we have the following result, which is crucial to our analysis.

Observation 2. *Each rank function on a given tree being equally likely is equivalent to the statement: For each internal node v, each shuffle at v is equally likely and these shuffles are independent.*

Neutral Models for Ranked Tree Shapes

In this section, we discuss 2 classes of neutral models. First we discuss a slightly generalized version of the equal rates Markov (ERM) model (Mooers and Heard 1997). Then we introduce the constant relative probability (CRP) models, which are a neutral class of models that allow clades to evolve with different rates. The common theme between these 2 classes of models is that they both induce the uniform distribution on tree shuffles for each tree.

It will be convenient to discuss ROTs rather than ranked phylogenetic trees, for reasons discussed below.

Definition 3. An "oriented tree" is a finite rooted binary tree where the children of each internal node are labeled *left* and *right*, respectively. An ROT is an oriented tree with a rank function.

These trees are called "oriented" because they are oriented graphs, that is, the edges around each vertex have a fixed orientation. Oriented trees can also be called "planar" trees as they are equivalent to a depiction of a tree in the plane.

ROTs are convenient as they allow us to distinguish the children of each vertex without having to explicitly label species which may later become extinct. It might seem more natural to consider phylogenetic trees, that is, trees with leaf labels, by assigning a unique label for each new leaf and eliminating the label in case of extinction. For trees on n leaves evolving under a bifurcation and extinction model, however, we cannot guarantee that all such trees on n species will have the same leaf labels (due to random extinction events). On the other hand, by considering oriented trees, we can still distinguish between species because there is a unique path from the root to any leaf but do not need to worry about explicitly labeling leaves and what to do with extinct leaf labels.

We will call ranked trees without orientation or labeling "ranked tree shapes". The runs statistic operates on ranked tree shapes as it is independent of orientation and labeling of a tree. Therefore, for applications, we will be comparing the ranked tree shape distribution induced by our neutral models to the ranked tree shape given by reconstructed trees.

We now explain how the oriented trees are generated by an evolutionary model, by which we mean a "forward time ranked-oriented-tree-valued stochastic process". In particular, we consider birth–death processes where the transitions between trees involve either a bifurcation (e.g., speciation) or an extinction event. Call these "ranked-oriented-tree birth–death processes". The details of bifurcation (birth) and extinction (death) events are as follows. If there is a bifurcation event, the new branches descending from the bifurcation event are assigned *left* and *right*. If there is an extinction event, occurring at a leaf vertex, the leaf and its adjacent edge are deleted. The ancestor of the extinct leaf is now a degree-2 vertex. This vertex is suppressed by replacing it and its 2 adjacent edges by a single edge, with orientation inherited from the edge closer to the root. In this way, the ancestor still has a left and right child. The ranking of internal vertices is induced by the time ordering of their associated bifurcation events. If extinction events are not allowed, then call such a process a "pure-birth ranked-oriented-tree process".

We will now discuss 2 classes of ranked-oriented-tree birth–death stochastic processes as neutral models for bifurcation and extinction, the ERM models and the CRP models.

ERM Models

We define an ERM model to be a forward time ranked-oriented-tree birth–death process such that any new bifurcation or extinction event is equally likely to occur in any extant lineage. We will consider the projection of this process onto ranked tree shapes by forgetting the orientation of children at each internal node. Any model described in terms of rates is an ERM model if the bifurcation and extinction rates are equal between all lineages at any given time. This class of models includes the Yule (1924) model, the critical branching process model (Aldous and Popovic 2005), the constant rate birth and death process (Nee, May, and Harvey 1994), and the coalescent (Kingman 1982).

Note that the ERM class includes models that have bifurcation and extinction rates varying in an arbitrary fashion depending on time or the current state of the

process, such as global speciation and extinction rate variation due to global environmental conditions. Furthermore, it is also possible to incorporate models with random incomplete taxon sampling, which is equivalent to the deletion of k species uniformly at random from the complete tree. Namely, if the complete tree evolved under an ERM model, then we simply run the model longer with the probability of bifurcation set to zero and the extinction probability nonzero and uniform across taxa. This extended model is clearly still within the ERM class.

The ERM class also includes microevolutionary models such as the coalescent with arbitrary population size history. This very simple but important fact means that the tests for nonneutral diversification described in later sections are not "fooled" by ancestral population size variation, as are a number of other tests in the literature.

The crucial fact about these models which makes our statistical analysis possible is that ERM models give the uniform distribution on ROTs as well as the uniform distribution on rank functions given a tree shape. This fact is demonstrated in the appendix, but we give a simple version here. The starting point is that all ROTs are equiprobable under an ERM model. This can be seen by induction: say it is true for all trees with n tips. Each new tree is uniquely determined by choosing such a tree (with equal probability, by induction) and then choosing a tip to bifurcate (with equal probability, by the ERM assumption). This then gives a uniform distribution on trees with $n + 1$ tips. Extinction can be dealt with in a similar way, as described in the appendix.

The uniform distribution on ROTs yields the uniform distribution on rank functions conditioned on the oriented tree. Therefore, each orientation of a given tree shape gives the uniform distribution on rank functions; a weighted sum of uniform distributions is still uniform, so we have a uniform distribution on rank functions conditioned on the tree shape. We also show in the appendix that, for the case of pure-birth models, the ERM models are the only models for bifurcation and extinction which induce a uniform distribution on ROTs. These properties of the ERM models will be useful for the runs test as described below.

CRP Models

The motivation for the CRP models comes from considering the models on ranked trees which might emerge from nonselective diversification, perhaps based on physical or reproductive barriers. For example, assume that we could watch a set of species emerge via allopatric speciation, and the fundamental geographic barrier divides the land into 2 regions, A and B. These regions may differ in size or fecundity, so there may be some difference in the rate of diversification in A versus B. However, our neutral assumption for the CRP class is that the "relative" rate stays constant over time. In contrast, nonneutral models might dictate that a bifurcation in one region will shift the equilibrium such that further diversification in that region will become more likely ("bursting" diversification) or less likely ("refractory" diversification).

Again, for convenience, we work with ROTs so we may distinguish the 2 children of any bifurcation event. For each internal node, v (representing a bifurcation event), let L_v and R_v denote the "left" and "right" lineages descending from v (daughter subtrees of v).

A CRP model is a forward time pure-birth ranked-oriented-tree process together with a probability distribution P on the unit interval $[0, 1]$, where each internal node v has a real number, p_v, associated with it. Each new bifurcation occurring in the clade below v occurs in L_v with probability p_v and occurs in R_v with probability $1 - p_v$. For each new bifurcation event (internal node), v, choose the value p_v by an independent draw from P. Note that CRP models are Markov processes on trees equipped with probabilities p_v for every internal node v.

In other words, given a tree on n species, the CRP model chooses a leaf to bifurcate as follows. At the root ρ of T, we choose the left daughter tree L_ρ with probability p_ρ versus the right daughter tree R_ρ with probability $1-p_\rho$. Assume for the sake of description that we choose L_ρ, with root vertex v. Then, as before, the next bifurcation will happen in L_v with probability p_v and R_v with probability $(1 - p_v)$. Repeat in this manner until a leaf is reached. The bifurcation of that leaf makes a new internal node w, for which we draw p_w by an independent draw from P. For an example, see Figure 3.

Note that the CRP and ERM models are disjoint. Assume that we are watching a CRP tree grow such that at the stage with 2 leaves, the probability of picking 1 of each of the 2 leaves is equal. Then after the next event, we will have a tree with 3 leaves, which cannot all have the same probability of bifurcation. Indeed, if the 2 new leaves are to have equal probability of bifurcation, then they must have probability $1/4$ each, whereas the leaf that did not bifurcate in the second event has probability $1/2$ of bifurcation.

Given the CRP definition, it should not be too hard to believe that CRP models induce a uniform distribution on shuffles for a given oriented tree or tree shape.

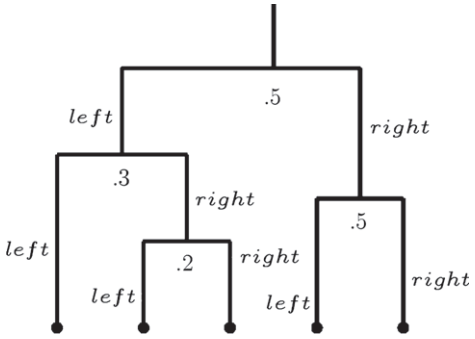

FIGURE 3. An oriented tree that evolved under the CRP model with the bifurcation probabilities p_v for each internal node v. Assuming that we see a speciation event next, the probability for the farthest left leaf to speciate next is $0.5 \times 0.3 = 0.15$. In contrast, under the ERM model, each leaf is equally likely to speciate with probability $1/5$.

Namely, assume that we are looking at lineages descending from an internal node v. Any shuffle with n descendants on the left side and m descendants on the right side of v will have probability $p_v^n(1-p_v)^m$. A complete argument showing that the CRP models give uniform distributions on shuffles is given in the appendix. The CRP may be viewed as a generalization of the stick-breaking models (Aldous 1995) from tree shapes to ranked trees, but this argument is omitted.

The CRP definition shows how tree balance and shuffle structure can be independent sources of information. Indeed, imagine that the probability mass of the distribution P is concentrated on the boundaries of the interval, so that p_v is always very close to 0 or 1. In such a case, the resulting trees will be very imbalanced (as there will always be a significant difference in relative rates of diversification between 2 sister clades), but as for any CRP model the shuffle distribution will be uniform. On the other hand, if P is concentrated near the middle of the interval, then the resulting trees will tend to be quite balanced.

As shown in the appendix, the ERM models are the only pure-birth models that induce a uniform distribution on ROTs. The ERM and the CRP models both induce a uniform distribution on rankings conditional on the tree topology. Stadler (2008) has determined the whole class of pure-birth models that induce a uniform distribution on rankings conditional on the tree topology. This class is defined by a set of conditions that are difficult to interpret in a biological sense, and only the ERM and the CRP models appear to have a simple biological interpretation.

Tests for Bursting Diversification based on Shuffles

In the previous section, we discussed the ERM and CRP models and observed that they induce the uniform distribution on tree shuffles. In this section, we describe a way of testing for deviation from the uniform distribution on tree shuffles and thus test for deviation from these neutral models. We emphasize that this can go beyond testing the coalescent or Yule models, which are typically considered to be the definition of neutrality. Indeed, rejection of the uniform distribution on shuffles rejects all the ERM and CRP models simultaneously. Although the focus of this section is to consider all the shuffles of a ranked tree at once, one can also consider a shuffle at a particular node as described in the Introduction.

There are several useful tools available to test whether a shuffle is likely to have come from the uniform distribution on shuffles. In fact, a number of classical statistical tests for equality of distributions (e.g., the runs test, the Mann–Whitney–Wilcoxon test) actually implement a test of deviation from the uniform distribution on shuffles. These tests work as follows: assume that we are given 2 sets of real samples $\{\ell_i\}_{i=1,\dots,m}$ and $\{r_j\}_{j=1,\dots,n}$ and would like to test the hypothesis that they are draws from the same distribution. To test, combine the draws and put the samples in increasing order (assume that all draws are distinct). This clearly gives a shuffle on symbols ℓ and r. If the draws are from identical distributions, then the induced distribution on shuffles will be uniform; if on the other hand symbols cluster together in the shuffle, there is some evidence that the draws are from unequal distributions.

One can then test deviation from the uniform distribution on shuffles in one of several ways. One way is to count the number of "runs." As described in the Introduction, a run is simply a sequence within the shuffle using only one symbol; the shuffle $\ell\ell rrr\ell$ has 3 runs. Let $X_{m,n}$ denote the number of runs under the uniform distribution on shuffles on m symbols of one type and n of another. The distribution of $X_{m,n}$ is classical (see, e.g., Hogg and Craig 1994):

$$\begin{aligned}\mathbb{P}\{X_{m,n}=2k+1\}&=\frac{\binom{m-1}{k}\binom{n-1}{k-1}+\binom{m-1}{k-1}\binom{n-1}{k}}{\binom{m+n}{m}},\\ \mathbb{P}\{X_{m,n}=2k\}&=\frac{2\binom{m-1}{k-1}\binom{n-1}{k-1}}{\binom{m+n}{m}}.\end{aligned} \tag{1}$$

Asymptotic results for the mean and variance are also known:

$$\begin{aligned}\mathbb{E}[X_{m,n}]&=\mu_{m,n}=2\frac{mn}{m+n}+1,\\ \mathrm{Var}[X_{m,n}]&=\frac{(\mu_{m,n}-1)(\mu_{m,n}-2)}{m+n-1}.\end{aligned} \tag{2}$$

The usual application of the runs test makes a shuffle from the 2 draws as described above, calculates the number of runs in the shuffle, and then uses the above-calculated probabilities to test deviation from the uniform distribution on shuffles. However, the same method can be applied in any situation to test deviation from the uniform distribution on shuffles. In the present case, we can use an analogous process to investigate tree shuffles.

As described in the Introduction, a tree shuffle simply assigns to each internal node of the tree a shuffle of the appropriate type, (m,n); from the previous section, we know these shuffles to be distributed uniformly and independently for a variety of neutral models. Using runs we can test whether a single shuffle is drawn from the uniform distribution, but some method is needed to combine this information across the internal nodes of the tree.

We chose to combine our data from each vertex by simply summing the number of runs across all the shuffles in the corresponding tree shuffle. Let $\mathcal{R}(T)$ denote the number of runs for T, a tree shape, an oriented tree, or a phylogenetic tree. The distribution of $\mathcal{R}(T)$ under the assumption that each tree shuffle is equally likely can be calculated recursively as shown in the next 2 sections.

We calculate 2 distinct null distributions for $\mathcal{R}$. First, we condition on the observed tree shape T and calculate

the distribution of $\mathcal{R}(T)$ under the uniform distribution of shuffles on T, leading to what we call p_{cond} below. Second, we calculate the distribution of $\mathcal{R}(T)$, where T is drawn from the uniform distribution on ROTs, which is the distribution corresponding to the ERM model. As shown below, the uniform distribution on ROTs induces the uniform distribution on shuffles on any single tree, but the converse is not true.

Note that there are a number of alternative ways to test deviation from the uniform distribution on shuffles. First, we have made one choice—namely, summation—concerning how the statistics for each shuffle are combined. One certainly could use an alternative method, potentially including weights. Second, there are other statistics, such as Mann–Whitney–Wilcoxon, that which could be used in place of the runs statistic. The advantage of summation is that it results in simple formulas, and the advantage of the runs statistic is that it is easy to interpret. We have not tested any alternate formulations.

p_{cond}: Testing conditioning on an observed tree shape. We first condition on the observed tree, defining a p value which we will call the "conditioned runs p value", or p_{cond}. We do so because the imbalance of a given tree shape has a substantial influence on the range of the total number of runs. Indeed, consider a given internal node with an (m, n) shuffle: that is, one daughter subtree has m internal nodes and the other has n internal nodes. If $m = n$, then the maximum number of runs at that internal node is equal to $2n$. If instead $m > n$, then the maximum number of runs is $2n+1$. Thus, the maximum number of runs at an internal node is bounded above by a function of the size of the smallest descendant subtree; consequently imbalanced trees will have a smaller upper bound for the total number of runs compared with balanced trees. Informally, when selecting a tree from some distribution on tree shapes, imbalanced trees will "tend" to have fewer runs than balanced trees. However, by conditioning on a given tree shape, we eliminate the contribution of tree shape to the distribution on the number of runs and focus only on timing information.

As shown in the appendix, the ERM and CRP models each induce the uniform distribution on shuffles conditioned on a tree shape (Corollaries A2 and A6 and Proposition A8). For a tree with 1 leaf, we have $\mathbb{P}\{\mathcal{R}(T)=0\} = 1$. For a tree with 2 leaves, we also have $\mathbb{P}\{\mathcal{R}(T) = 0\} = 1$ (the 2 daughter subtrees have no internal nodes).

Assume that a random tree T from a distribution inducing the uniform distribution on shuffles is composed of 2 ranked subtrees L and R of size m and n, respectively. Then, we have

$$\mathbb{P}\{\mathcal{R}(T) = k\} = \sum_{i=0}^{k} \mathbb{P}\{X_{m-1,n-1} = i\} \times \sum_{j=0}^{k-i} \mathbb{P}\{\mathcal{R}(L) = j\}\mathbb{P}\{\mathcal{R}(R) = k - i - j\}. \quad (3)$$

It is shown in the appendix that this distribution can be calculated recursively on a tree with n leaves in time $O(n^3 \log^2 n)$, with small constant. Thus, it is practical to obtain a p value for $\mathcal{R}(T)$ analytically. If n is ever so large that this computation is prohibitively expensive, then simulation may be used to efficiently approximate the cumulative distribution function (CDF) of $\mathcal{R}(T)$ and thus the p value.

We define the conditioned runs p value $p_{\text{cond}}(S)$ for a ranked tree shape S as

$$p_{\text{cond}}(S) = \sum_{i=0}^{\mathcal{R}(S)-1} \mathbb{P}\{\mathcal{R}(T) = i\} + \frac{1}{2}\mathbb{P}\{\mathcal{R}(T) = \mathcal{R}(S)\},$$

where $\mathbb{P}\{\mathcal{R}(T) = k\}$ is computed from equation (3). We use this quantity rather than $\mathbb{P}\{\mathcal{R}(T) < \mathcal{R}(S)\}$ or the classical $\mathbb{P}\{\mathcal{R}(T) \leq \mathcal{R}(S)\}$, as these alternative formulations lead to uninformative extreme p values for small trees.

p_{unif}: Testing the uniform distribution on ROTs. Now for the second case, we define the "uniform runs p value" p_{unif}, assuming that we want to test a model such that each ROT is equally likely. This includes the ERM models, and in the case of pure-birth models, this is exactly the set of the ERM models (Proposition A7). Let $\mathcal{R}(n)$ be the random variable "runs of a tree with n leaves," where the tree is drawn from the uniform distribution on ROTs. The distribution of $\mathcal{R}(n)$ can again be obtained recursively. Note that for a uniform ranked tree on n leaves, the probability that one daughter tree has size r and the other daughter tree has size $n - r$ is $1/(n-1)$ for all r. Thus,

$$\mathbb{P}\{\mathcal{R}(n) = k\} = \frac{1}{n-1} \sum_{r=1}^{n-1} \sum_{i=1}^{k} \mathbb{P}\{X_{r-1,n-r-1} = i\} \times \sum_{j=0}^{k-i} \mathbb{P}\{\mathcal{R}(r) = j\}\mathbb{P}\{\mathcal{R}(n-r) = k - i - j\}. \quad (4)$$

The complexity for recursively calculating the distribution of runs for trees with n leaves is $O(n^4 \log^2 n)$, by an argument analogous to that for equation (3).

We define the uniform runs p value $p_{\text{unif}}(S)$ for a ranked tree shape S as

$$p_{\text{unif}}(S) = \sum_{i=0}^{\mathcal{R}(S)-1} \mathbb{P}\{\mathcal{R}(n) = i\} + \frac{1}{2}\mathbb{P}\{\mathcal{R}(n) = \mathcal{R}(S)\},$$

where $\mathbb{P}\{\mathcal{R}(n) = k\}$ is computed from equation (3).

A ranked tree T that is more balanced than Yule model will tend to have $p_{\text{unif}}(T) > p_{\text{cond}}(T)$, whereas a tree less balanced than Yule model will tend to have the opposite relation. The underlying reason is that balanced ranked trees with shuffles drawn from the

uniform distribution tend to have more total runs than imbalanced ones; indeed, by equation (2), the expectation of $X_{m,k-m}$ is maximized when $m = \lfloor k/2 \rfloor$, that is, the perfectly balanced case. A maximally imbalanced vertex, on the other hand, can have only one run. These statements suggest the trend $p_{\text{unif}}(T) > p_{\text{cond}}(T)$ for balanced trees and $p_{\text{unif}}(T) < p_{\text{cond}}(T)$ for imbalanced trees because p_{unif} compares the given ranked tree to ranked trees with the Yule model distribution on tree shapes, whereas p_{cond} compares it to ranked trees with the same tree shape. It might be possible to make these statements more formal by fixing a notion of balance, but we do not pursue that here.

A python package for computing p_{cond} and p_{unif} is available at http://www.tb.ethz.ch/people/tstadler. For a collection of trees (e.g., a sample from the Bayesian posterior), the individual p values can be averaged. In addition to the calculation of the runs statistic and the p value for the whole tree, the package can calculate the runs statistic and p value for each interior vertex of a single tree. This feature may be useful for biologists looking for signals of a key innovation.

Shuffles in the Bayesian Setting

In our work up to now, we have assumed that the correct tree and diversification timing information are known. This assumption is not realistic for a number of data sets. For example, below we apply our methodology to a sample of HCVs, which probably do not have enough sequence divergence to perfectly reconstruct a phylogenetic tree with timing information.

One way of working with such data sets is to take a Bayesian approach, where rather than a single tree one gets a posterior distribution on trees. For each single tree, one can compute the p value of the total runs statistic, either conditioning on the topology or assuming a uniform distribution on ranked tree shapes. We then simply take the average of the p values thus computed for each tree. The average of p values in this case is a simple type of posterior predictive p value (Meng 1994; Ree 2005). As such, it is not exactly uniformly distributed under the neutral model as a proper p value should be, although the average does share many of the characteristics of a classical p value.

Runs and Neutrality

Here we note that the runs statistic can be used to test the coalescent in the presence of ancestral population size variation. Tests of neutrality in the presence of historical population size variation are of particular recent importance as new coalescent-based methods are in use to infer population size history in a Bayesian framework (Drummond et al. 2005; Opgen-Rhein et al. 2005). If these methods are to be used on a given set of sequences, it is important to test the central assumption of the methods, namely, that the sequences have a genealogy that can be accurately described using the coalescent with arbitrary population size history.

Unfortunately, classical statistics such as the D statistics of Tajima (1989) and Fu and Li (1993) confound ancestral population size changes and nonneutral evolution. One solution to this problem is to investigate the Bayesian posterior on phylogenetic trees for evidence of nonneutral evolution rather than using the sequence information directly. This has been done by Drummond and Suchard (2008), who use a posterior predictive p value approach. Here we simply point out that, as described above, the coalescent with arbitrary population size history is an ERM model and thus will induce the uniform distribution on ranked phylogenetic trees; therefore, by rejecting the ERM class we reject a general coalescent model. We will apply this fact below in the example application to HCV data.

Generalization for Nonbinary Trees

Polytomies (i.e., nonbinary splits) are common in reconstructed phylogenetic trees. Some polytomies are certainly due to a lack of information to resolve the splits; however, it has been argued that molecular- and species-level polytomies actually exist (Jackman et al. 1999; Slowinski 2001). The methodology described in this paper can be extended to trees with "hard" polytomies, that is, cases of multiple divergence which are essentially simultaneous in evolutionary time.

The new ingredient needed is the "multiple runs distribution," that is, the analog of equation (1) for shuffles on more than 2 symbols. Such distributions are described in David and Barton (1962). Using these distributions, the probability of a shuffle consisting of symbols from the k daughter trees can be found for a shuffle at a nonbifurcating split v in the same recursive manner.

Example Applications

In this section, we describe 2 applications of the methods in this paper. First, we apply the methods to relaxed clock phylogenetic trees of the HCV. The p_{cond} values for this data set show some bias toward LSB diversification. However, for certain data sets, p_{unif} clearly rejects any ERM model such as the coalescent with varying population size. The second application is to phylogenetic trees for ants, whose timing information was reconstructed through fossils and the r8s (Sanderson 2003) rates smoothing program. These ant trees do not show any evidence of LSB evolution, despite some interesting history in terms of diversification rates.

Our HCV data are derived from 5 independent studies in the literature. The first data set, "China," came from a survey of HCV in China (Lu et al. 2005). From this survey, we used 132 sequences of the E1 region. The second data set, "Donlin," consists of samples from 8 clinical centers in the United States (Donlin et al. 2007). From this study, we used the 48 sequences that were sampled in the year 2002. The third, "Egypt," is an alignment of E1 sequences from a survey of 71 HCV-infected individuals in Egypt (Ray et al. 2000). The

fourth alignment, "Henn," came from those HCV sequences with accession numbers EU155213–EU155344 which were sampled in the year 2006 (a total of 35 sequences; each from a distinct patient in the United States). "Timm," the fifth data set, supplies 71 sequences from a study of HCV-positive subjects in Boston (Timm et al. 2007). All sequences were downloaded from the Los Alamos HCV sequence database (Kuiken et al. 2005).

Phylogenetics in HCV is typically done on the core/envelope (CE) region, which codes for the nucleocapsid and envelope glycoproteins, or the nonstructural 5 (NS5) region, which codes for an interferon-resisting protein and RNA polymerase. We also restricted the alignments to these regions. Specifically, the study by Timm et al. (2007) sequenced a substantial segment including the NS5 region but not the CE region, so we restricted our study to the NS5 region. The studies by Donlin et al. and Henn et al. sequenced the complete genome, so we cut out the CE and the NS5 regions of the alignments and analyzed them separately. They are labeled as such below: for example, "Donlin CE" means the Donlin data set restricted to the CE region.

In order to avoid confounding temporal information with molecular rate variation, we applied the relaxed clock model of Drummond et al. (2006) as implemented in the BEASTv1.4 suite of computer programs (Drummond and Rambaut 2007). We chose uncorrelated lognormally distributed local clocks, the Hasegawa–Kishino–Yano model, and 4 categories of gamma rate parameters in the gamma + invariant sites model of sequence evolution. We performed separate analyses using both the constant population size and the exponential growth coalescent priors. All other parameters were left as default; the corresponding BEAST XML input files are available from the authors upon request.

The Markov chain Monte Carlo (MCMC) chain for these analyses was run for 10–100 million generations and convergence to stationarity was checked with the BEAST program Tracer. For each model parameter, the minimum effective sample size was at least 100, with most being significantly greater. The first 10% of the run was removed, and 100 trees were taken from the tree log file, equally spaced along the run of the MCMC chain. We interpret these trees as being independent samples from the posterior. As a check, the analysis was rerun with an empty alignment; no significant deviation from the uniform distribution on shuffles was detected (results not shown).

After running the BEAST analysis, we calculated expected p values for each data set using our statistics. First, we calculated p_{cond}, the p value of the number of runs conditioning on tree shape, calculated via equation (3). This is the p value of the total number of runs observed on the tree compared with the number of runs for a tree of the same shape where the shuffles are drawn from the uniform distribution. By thus conditioning on a tree topology, we consider the deviation of just the timing information from that of an ERM or a CRP model, rather than deviation of timing and topological information. Next we calculated p_{unif}, which is the p value of the observed number of runs under the assumption of the uniform distribution on ROTs. This p value tests ERM models, such as the coalescent with arbitrary population size history. Last we calculated a p value for the uniform distribution on ROTs using I_c, Colless' imbalance statistic (Colless 1982). I_c is calculated as follows: for each internal node v, let i_v be the difference of the number of leaves of the 2 daughter trees of v. Now I_c is simply the sum of the i_v for all internal nodes v. We included it for comparison, as it is a statistic that only includes tree shape and not timing information. The results are displayed in Table 1.

First, note that p_{Coll} is always less than one-half for these alignments. One such p value being less than one-half does not have any statistical significance; however, Table 2 shows that the combination of all such p values is highly statistically significant, meaning that the trees are more imbalanced than would be expected under neutrality. Thus, because imbalanced trees tend to have fewer runs than balanced trees (see the section introducing p_{cond}), one would expect that p_{cond} would be greater than p_{unif}; this is indeed the case here. Second, note that for all data sets except for Henn NS5, p_{cond} is quite a bit greater than p_{unif}. In those cases, the total number of runs is governed primarily by tree shape rather than by timing information. By looking at the table, clearly it is those cases where p_{cond} is much greater than p_{unif}, where p_{Coll} is small. On the other hand, for the Henn NS5 data set, p_{Coll} is large and p_{cond} is not too far from p_{unif}. Although the results for the CE and NS5 regions are relatively similar for the Donlin data set, the results for the same regions are quite different for the Henn data set. We do not have a clear explanation for why this would be.

We also note that the China and Egypt data sets, which focused on broad geographic sampling, appear to have more deviation from neutrality in terms of shape than clinic-based studies in the United States. One reason for this might be that the China and Egypt studies had enough sequence divergence to pull the observed distribution away from the prior in our BEAST analysis. Another possibility is that something about the sampling biased the tree shape away from neutrality.

In any case, it is clear from looking at Table 1 that for the China, Egypt, and Henn NS5 data sets, the observed trees deviate significantly from the distribution given by the arbitrary population size coalescent. This is topical as it is common practice to use the coalescent to estimate viral population size history, for example, the paper by Opgen-Rhein et al. (2005), which uses the Egypt data set. Other papers that use the coalescent to estimate population size history include Drummond et al. (2005) and Minin et al. (2008). This assumption goes untested as no methods were available at the time to test the coalescent in the presence of ancestral population size changes; it would be interesting to know how this would impact the historical population size estimates in these papers.

One might also take as a null hypothesis "HCV evolves according to some ERM or CRP model" and find a p value for that statement given this collection of independent studies. To do so, we combine p values as described below with the idea that although a single tree with a p value of around 0.25 has very little statistical significance, the fact that 5 independent studies give p values around that value is significant.

Combining p values is a common procedure in such situations with independent tests. See Loughin (2004) for a comparison of several methods. These 5 data sets are from separate studies; thus, we assume that the p values are independent. Here, we use 3 different methods of combining p values. They are the Anderson–Darling (A-D) test for uniformity (Marsaglia G. and Marsaglia J.C.W. 2004), sums of uniform quantiles, and sums of normal quantiles (Loughin 2004). As noted above, we obtained different results when choosing the CE region versus the NS5 region for building trees in the Donlin and Henn data sets; thus, for every combining method we made one choice of region for both Donlin and Henn and ran the analysis. The results are shown in Table 2.

We now describe the methods of combining p values. The A-D test uses a test statistic based on the difference between the empirical CDF and the CDF of the proposed distribution, in this case the uniform distribution as our samples are p values. This method is often used to test pseudorandom number generators, as in Marsaglia's set of DieHard tests (Marsaglia and Tsang 2002). The distribution of the A-D statistic is computed by simulation in this case, to find the p value for our combined sample of 5 numbers. The latter 2 tests involve choosing a distribution and summing the quantiles corresponding to each p value (see Loughin 2004).

All choices and tests reject the ERM model at the 5% level (seen in the columns labeled p_{unif} and p_{Coll}). The evidence from p_{cond}, conditioning on tree shape (and considering timing information only), is not so strong. In most cases, summing the quantiles of the uniform distribution is the most powerful test. As detailed in Loughin (2004), the uniform quantile method of p value combination is more powerful than the normal (and many others) against such cases where evidence against the hypothesis is weak and spread evenly among the contributing p values. The uniform test rejects the ERM and CRP models at the 5% level using p_{cond} for both the CE and the NS5 data when the exponential prior is used.

We would like to note that these results were found despite the fact that the coalescent was used as a prior in the Bayesian phylogenetic analysis. That is, if any bias could be expected in the trees, it would be toward a coalescent prior and a uniform distribution on shuffles. By looking at Tables 1 and 2, it is clear that the prior has a significant influence: indeed, just the difference between a constant population size and an exponentially increasing population size prior is enough to significantly change the timing and shape of the tree. Thus, we believe that our results form an upper bound for the actual statistics of the HCV lineages.

For the second application, we investigated 2 different trees of ant taxa. The first tree (149 taxa) is that of Moreau et al. (2006), showing the diversification of the major ant lineages. The timing information in this tree is quite remarkable, in that the corresponding LTT plot shows a substantial increase in diversification rate during the Late Cretaceous to Early Eocene, which corresponds to the rise of angiosperms (flowering plants). Given the tools at our disposal, one might wonder if this increase in diversification rates affected all lineages equally or if it occurred in lineage-specific bursts. The second ant tree we investigated was that of *Pheidole*, a "hyperdiverse" ant genus. *Pheidole* is almost certainly monophyletic and yet comprises about 9.5% of the ant species in the world, according to latest estimates (Moreau 2008). Moreau has recently reconstructed a phylogeny of this genus (171 taxa) which we have analyzed along with the tree of the ant lineages in general. Both trees were reconstructed via maximum likelihood and then made ultrametric using the penalized likelihood method of the r8s rates smoothing program (Sanderson 2003).

In Figure 4, we show a plot of the internal nodes of each tree. The x-coordinate in the plot is the number of taxa below an internal node, and the y-axis is the p value of the number of runs in the shuffle statistic. As can be seen, there is no clear correlation between the number of taxa below an internal node and the shuffle statistic, and at no stage does diversification appear to be consistently

TABLE 1. Expected p values of the number of runs in the posterior for a Bayesian analysis of HCV data

Data set	Const p_{cond}	Exp p_{cond}	Const p_{unif}	Exp p_{unif}	Const p_{Coll}	Exp p_{Coll}
China	0.102	0.0778	0.00122	0.000454	0.00059	0.000377
Donlin CE	0.32	0.193	0.14	0.13	0.094	0.307
Donlin NS5	0.365	0.455	0.156	0.193	0.106	0.122
Egypt	0.344	0.322	0.03	0.0131	0.00391	0.00162
Henn CE	0.224	0.243	0.106	0.122	0.117	0.13
Henn NS5	0.0617	0.0388	0.0477	0.0363	0.324	0.334
Timm NS5	0.576	0.464	0.389	0.271	0.154	0.0879

Notes: Each row represents one data set. "Const" means that the BEAST analysis used a constant population size coalescent prior, and "Exp" denotes the use of an exponentially increasing population size coalescent prior. We display results using 3 methods of testing as described in the text: the conditioned runs p value p_{cond}, the uniform runs p value p_{unif}, and the p value for the Colless index p_{Coll} testing against an ERM model. The p_{unif} columns reject ERM models for certain data sets, as do the p_{Coll} columns.

TABLE 2. Combined runs p values for the 5 studies of HCV data

Data set	Const p_{cond}	Exp p_{cond}	Const p_{unif}	Exp p_{unif}	Const p_{Coll}	Exp p_{Coll}
A-D CE	0.781	0.9013	0.99803	0.999479	0.999939	0.999875
A-D NS5	0.829	0.887	0.99896	0.999735	0.999614	0.999857
Uniform CE	0.076	0.0311	0.00112	0.000397	7.4×10^{-5}	0.000359
Uniform NS5	0.0524	0.0384	0.00079	0.000277	0.000592	0.000387
Normal CE	0.0666	0.0434	0.00021	2×10^{-5}	3×10^{-5}	$<1 \times 10^{-6}$
Normal NS5	0.095	0.0582	0.00953	0.0068	0.0103	0.0208

Notes: We combine using the A-D test for uniformity, sums of uniform quantiles (labeled Uniform), and sums of normal quantiles (labeled Normal). The rows are also labeled CE or NS5 based on if we used the CE or the NS5 region for phylogenetic analysis in the Donlin and Henn data sets.

bursting or refractory in a lineage-specific sense. We can also compute the p value of the total number of runs across the tree: for tree (a) this is about 0.9052 and for tree (b) this is about 0.6718. Thus, for these 2 ant trees we do not see any significant evidence of LSB or refractory diversification. This analysis forms an interesting counterpoint to the LTT results for the ants, which shows an overall increase in diversification rate during the Late Cretaceous to Early Eocene across the entire tree.

CONCLUSIONS

We have developed a framework that allows testing for nonneutral diversification timing. Our work consists of 3 main components: first, a simple, recursive way of quantifying the relative timing information on a phylogenetic tree; second, 2 classes of neutral models on trees with relative timing information; and third, a summary statistic that allows comparison of reconstructed trees to these neutral models. In our methodology, timing information is considered relative to sister taxa and considered in the context of the tree, which may make it a valuable complementary method to LTT plots. We compute the significance of the deviation of relative timing information from a neutral model analytically using a simple method drawn from classical statistics.

This method was conceived for the macroevolutionary case in order to find historical evolutionary patterns requiring explanation. However, it is also quite applicable in the microevolutionary case, where it can test neutrality in the presence of historical population size variation. This is particularly relevant as methods are now available to describe historical population size under a coalescent assumption.

We emphasize that our methodology can go beyond testing for deviation from the coalescent or the Yule models, which are usually the entire class of "neutral" models considered. Indeed, because any ERM or CRP model induces the uniform distribution on shuffles, deviation from this distribution is evidence to reject any model in the ERM or CRP classes.

However, sometimes one may wish to test only a more restricted set of models, such as only the ERM models (which include the coalescent with arbitrary population size history) and not the ERM and CRP models together. By testing a more restricted class of models, a particular data set will be more likely to fall outside the chosen class. For example, in the application of our methods to HCV data above, the data show some weak evidence of not coming from an ERM or a CRP model. However, if one tests for conformity to the ERM class (again, including the coalescent with arbitrary

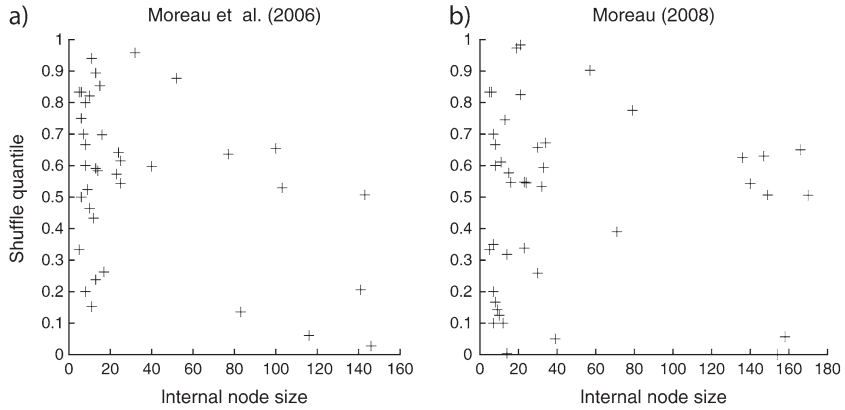

FIGURE 4. The distribution of the runs statistic for the internal nodes in 2 trees of ant taxa. Each point in each plot represents an internal node in the corresponding tree; the x-axis gives the number of taxa below the internal node, and the y-axis gives the p value of that internal node in terms of the runs statistic. The plot on the left is for the tree of Moreau et al. (2006), and the plot on the right is for a tree of *Pheidole* appearing in Moreau (2008). These 2 trees do not appear to consistently show either LSB or refractory diversification.

population size history), one obtains rejection at the 5% level.

We recall that our method uses "relative" timing information rather than actual branch lengths. In some ways this is an advantage. In a microevolutionary setting, this means that the corresponding tests are invariant to changes in ancestral population size, and thus our test for neutrality is not "fooled" by ancestral population size variation. In a macroevolutionary setting, the statistics are robust to branch length estimation error over long timescales. Such estimations are known to be difficult (Kimura 1981). We note further that from a modeling perspective it is possible to specify a probability distribution on ranked phylogenetic trees without specifying a particular distribution on branch lengths. This flexibility means that it may be possible to reject many models at once as described above.

Nevertheless, it may be useful at some future stage to combine topology and continuous branch length information, rather than the discretized version considered here. However, quantifying the shape of such objects appears to be challenging as the relevant geometry is quite intricate (Billera et al. 2001; Moulton and Steel 2004). In contrast, by discretization to ranked trees we obtain a purely combinatorial object.

We close by noting that although various techniques for reconstructing phylogenetic trees with timing information have been present for many years, these methods are currently seeing an intense period of development and will only improve. With this improvement we expect to see an increase in the number of trees present in the literature with interesting patterns of diversification timing due to adaptive radiation or other factors. We hope that our technique will prove to be a useful analytical tool for these future investigations, not only for finding interesting diversification patterns but also for testing potential biases of timing reconstruction methods.

Supplementary Data

Supplementary material can be found at http://sysbio.oxfordjournals.org/.

Funding

Ph.D. scholarship from the Deutsche Forschungsgemeinschaft (to T.S.); Miller Institute for Basic Research in Science at the University of California, Berkeley (to F.A.M.).

Acknowledgments

The idea of using ranked trees in this context was suggested by John Wakeley. The authors would like to thank Alexei Drummond, Arne Mooers, Allen Rodrigo, and Dennis Wong for helpful discussions. Mike Steel made several important suggestions, including the recursive calculation of p values for the shuffles. Alexei Drummond gave very valuable advice on running BEAST. John Huelsenbeck generously allowed us to run BEAST on his cluster. Corrie S. Moreau supplied trees for analysis and gave many very helpful comments on the manuscript. The manuscript was much improved by comments from Jack Sullivan, associate editor Cécile Ané, and 3 anonymous reviewers.

References

Aldous D. 1995. Probability distributions on cladograms. In: Aldous D., Pemantle R., editors. Random discrete structures. Berlin: Springer. p. 1–18.

Aldous D., Diaconis P. 1986. Shuffling cards and stopping times. Am. Math. Mon. 93:333–348.

Aldous D., Popovic L. 2005. A critical branching process model for biodiversity. Adv. Appl. Prob. 37:1094–1115.

Billera L.J., Holmes S.P., Vogtmann K. 2001. Geometry of the space of phylogenetic trees. Adv. Appl. Math. 27:733–767.

Colless D.H. 1982. Phylogenetics: the theory and practice of phylogenetic systematics. Syst. Zool. 31:100–104.

David F.N., Barton D.E. 1962. Combinatorial chance. New York: Hafner.

Donlin M., Cannon N., Yao E., Li J., Wahed A., Taylor M., Belle S., Di Bisceglie A., Aurora R., Tavis J. 2007. Pretreatment sequence diversity differences in the full-length hepatitis C virus open reading frame correlate with early response to therapy. J. Virol. 81: 8211–8224.

Drummond A., Ho S., Phillips M., Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond A., Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Drummond A., Suchard M. 2008. Fully Bayesian tests of neutrality using genealogical summary statistics. BMC Genet. 9:68.

Drummond A.J., Rambaut A., Shapiro B., Pybus O.G. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22:1185–1192.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein J. 1988. Phylogenies from molecular sequences: inference and reliability. Ann. Rev. Genet. 22:521–565.

Ford D.J. 2006. Probabilities on cladograms: introduction to the alpha model [Ph.D. thesis]. [Stanford (CA)]: Stanford University.

Fu Y.X., Li W.H. 1993. Statistical tests of neutrality of mutations. Genetics. 133:693–709.

Gillespie J. 1984. The molecular clock may be an episodic clock. Proc. Natl. Acad. Sci. USA. 81:8009–8013.

Harmon L., Schulte J., Larson A., Losos J. 2003. Tempo and mode of evolutionary radiation in iguanian lizards. Science. 301:961–964.

Hogg R.V., Craig A. 1994. Introduction to mathematical statistics. 5th ed. Upper Saddle River (NJ): Prentice Hall.

Huelsenbeck J., Larget B., Swofford D. 2000. A compound poisson process for relaxing the molecular clock. Genetics. 154:1879–1892.

Jackman T.R., Larson A., de Queiroz K., Losos J.B. 1999. Phylogenetic relationships and tempo of early diversification in Anolis lizards. Syst. Biol. 48:254–285.

Kimura M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA. 78: 454–458.

Kingman J.F.C. 1982. On the genealogy of large populations. J. Appl. Prob. 19A:27–43.

Kuiken C., Yusim K., Boykin L., Richardson R. 2005. The Los Alamos HCV sequence database. Bioinformatics. 21:379–384.

Losos J., Adler F. 1995. Stumped by trees—a generalized null model for patterns of organismal diversity. Am. Nat. 145:329–342.

Loughin T. 2004. A systematic comparison of methods for combining p-values from independent tests. Comput. Stat. Data Anal. 47: 467–485.

Lu L., Nakano T., He Y., Fu Y., Hagedorn C.H., Robertson B.H. 2005. Hepatitis C virus genotype distribution in China: predominance of closely related subtype 1b isolates and existence of new genotype 6 variants. J. Med. Virol. 75:538–549.

Maddison W., Midford P., Otto S. 2007. Estimating a binary character's effect on speciation and extinction. Syst. Biol. 56:701–710.
Marsaglia G., Marsaglia J.C.W. 2004. Evaluating the Anderson-Darling distribution. J. Stat. Softw. 9:1–5.
Marsaglia G., Tsang W.W. 2002. Some difficult-to-pass tests of randomness. J. Stat. Softw. 7:1–8.
Meng X.-L. 1994. Posterior predictive *p*-values. Ann. Stat. 22:1142–1160.
Minin V., Bloomquist E., Suchard M. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. 25:1459.
Mooers A., Harmon L.J., Blum M.G.B., Wong D.H.J., Heard S. 2007. Some models of phylogenetic tree shape. In: Gascuel O., Steel M., editors. Reconstructing evolution: new mathematical and computational advances. Oxford: Oxford University Press. p. 149–170.
Mooers A.O., Heard S.B. 1997. Evolutionary process from phylogenetic tree shape. Q. Rev. Biol. 72:31–54.
Moreau C., Bell C., Vila R., Archibald S., Pierce N. 2006. Phylogeny of the ants: diversification in the age of angiosperms. Science. 312: 101–104.
Moreau C.S. 2008. Unraveling the evolutionary history of the hyperdiverse ant genus *Pheidole*. Mol. Phylogenet. Evol. 48:224–239.
Moulton V., Steel M. 2004. Peeling phylogenetic 'oranges'. Adv. Appl. Math. 33:710–727.
Nee S., Holmes E., May R., Harvey P. 1994. Extinction rates can be estimated from molecular phylogenies. Philos. Trans. R. Soc. B. 344: 77–82.
Nee S., May R., Harvey P. 1994. The reconstructed evolutionary process. Philos. Trans. R. Soc. B. 344:305–311.
Opgen-Rhein R., Fahrmeir L., Strimmer K. 2005. Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. BMC Evol. Biol. 5:6.
Page R.D.M. 1991. Random dendrograms and null hypotheses in cladistic biogeography. Syst. Zool. 40:54–62.
Pagel M. 1997. Inferring evolutionary processes from phylogenies. Zool. Scripta. 26:331–348.
Paradis E. 1997. Assessing temporal variations in diversification rates from phylogenies: estimation and hypothesis testing. Proc. R. Soc. B. 264:1141–1147.
Paradis E. 1998. Detecting shifts in diversification rates without fossils. Am. Nat. 152:176–187.
Pybus O.G., Harvey P.H. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proc. R. Soc. B. 267:2267–2272.
Rabosky D. 2006. Likelihood methods for detecting temporal shifts in diversification rates. Evolution. 60:1152–1164.
Rabosky D., Lovette I. 2008a. Density-dependent diversification in North American wood warblers. Proc. R. Soc. B. 275:2363–2371.
Rabosky D., Lovette I. 2008b. Explosive evolutionary radiations: decreasing speciation or increasing extinction through time? Evolution. 62:1866–1875.
Ray S.C., Arthur R.R., Carella A., Bukh J., Thomas D.L. 2000. Genetic epidemiology of hepatitis C virus throughout Egypt. J. Infect. Dis. 182:698–707.
Ree R. 2005. Detecting the historical signature of key innovations using stochastic models of character evolution and cladogenesis. Evolution. 59:257–265.
Ricklefs R. 2007. Estimating diversification rates from phylogenetic information. Trends Ecol. Evol. 22:601–610.
Sanderson M. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics. 19:301–302.
Schluter D. 2000. The ecology of adaptive radiation. Oxford: Oxford University Press.
Semple C., Steel M. 2003. Phylogenetics. Oxford: Oxford University Press (Oxford lecture series in mathematics and its applications; vol. 24).
Slowinski J.B. 2001. Molecular polytomies. Mol. Phylogenet. Evol. 19:114–120.
Stadler T. 2008. Evolving trees—models for speciation and extinction in phylogenetics [Ph.D. thesis]. [Munich (Germany)]: Technische Universität München.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585–595.
Timm J., Li B., Daniels M., Bhattacharya T., Reyor L., Allgaier R., Kuntzen T., Fischer W., Nolan B., Duncan J., Schulze zur Wiesch J., Kim A.Y., Frahm N., Brander C., Chung R.T., Lauer G.M., Korber B.T., Allen T.M. 2007. Human leukocyte antigen-associated sequence polymorphisms in hepatitis C virus reveal reproducible immune responses and constraints on viral evolution. Hepatology. 46:339–349.
Weir J. 2006. Divergent timing and patterns of species accumulation in lowland and highland neotropical birds. Evolution. 60:842–855.
Yule G.U. 1924. A mathematical theory of evolution: based on the conclusions of Dr. J.C. Willis. Philos. Trans. R. Soc. Lond Ser. B. 213: 21–87.
Zink R., Slowinski J. 1995. Evidence from molecular systematics for decreased avian diversification in the Pleistocene epoch. Proc. Natl. Acad. Sci. USA. 92:5832–5835.

First submitted 14 March 2008; reviews returned 30 May 2008; final acceptance 10 January 2009
Associate Editor: Cécile Ané

APPENDIX

ROTs, Phylogenetic Trees, and Tree Shapes

For the following proposition, we define a "symmetric vertex" to be one for which the unoriented shapes of the subtree below each child of this vertex are the same (isomorphic as tree shapes). A "big symmetric vertex" is a symmetric vertex with more than 2 leaves below.

Proposition A1. A uniform distribution on rank functions on a given oriented tree induces a uniform distribution on rank functions of its corresponding tree shape.

Proof. Let t be an oriented tree with n leaves and t' its corresponding tree shape. Let q denote the number of big symmetric vertices of t or t'. For $n=2$, which implies $q=0$, we have for the ranking on t' exactly $1=2^0$ ranking on t. Assume that the following claim is true for all oriented trees with less than n leaves: for each ranking on t' there are exactly 2^q rankings on t which are sent to it by the map which forgets orientation of vertices. This claim implies the proposition as then the number of rankings on t which get sent to a given ranking R on t' does not depend on the choice of R. The induction now breaks into 2 cases.

Case 1: Suppose the 2 children of the root branch point of t' are nonisomorphic tree shapes. They may therefore be distinguished from each other, and given a ranking on t' the shuffle at the root node of t is determined. Call the 2 child subtrees "left" and "right" with q_1 and q_2 being big symmetric vertices, respectively. By the inductive assumption, there are 2^{q_1} rankings for the left subtree of t and 2^{q_2} rankings for the right subtree which map to the corresponding rank function on the left and right subtree shapes. This gives $2^{q_1+q_2}$ total because there is no choice for the shuffle at the root branch point of t. This completes the induction as $q_1 + q_2$ is the number of big symmetric vertices of t'.

Case 2: Suppose that the 2 children of t' are isomorphic. Therefore, they may not be distinguished except

by the ranking. Therefore, the shuffle at the root branch point of t is only determined up to swapping the left and right subtrees. After this choice the 2 subtrees are distinguished: which subtree of t' is "left" and which is "right" is determined by the shuffle. The rest of the argument proceeds as before, except that this time there are $2^{q_1+q_2+1}$ rank functions on t which map to the given rank function on t', and $q_1 + q_2 + 1$ is the number of big symmetric branch points of t'.

The result now follows by induction. □

Corollary A2. If a probability function on ROTs is uniform on rank functions conditioned on an oriented tree, then it is also uniform on rank functions of an (unoriented) tree shape when conditioned on that (unoriented) tree shape.

Proof. This follows from the previous proposition because the resulting mixture of uniform distributions on rank functions on t' (one for each oriented tree t with shape t') is also uniform. □

This corollary allows us to apply our rank tests to trees which are given without orientation: ranked tree shapes.

The ERM Model

In this section, we prove the properties of the ERM models mentioned in the main text. The following 2 facts are needed for the proof of Proposition A5.

Proposition A3. There are $(n-1)!$ ROTs on n leaves.

Proof. Proceed by induction on n; for $n=2$ the statement is obviously true, there is only one ROT. Suppose there are $(n-1)!$ ROTs with n leaves. For any tree on n leaves, there are n possibilities to attach a leaf which evolved at the nth bifurcation event. Attaching a leaf to any of the trees on n leaves gives us the set of ROTs on $n+1$ leaves. Thus, there will be $n(n-1)!=n!$ ROTs with $n+1$ leaves. □

Lemma A4. *Given an ROT with n leaves, there are $n(n+1)$ ways to add an additional leaf.*

Proof. First, decide which rank the new internal node will have, from 1 (earliest) to n (latest). If the new internal node has rank k, then there are k choices at that level for the edge to add it to, and then 2 choices for which side of this edge the new pendant leaf will sit. This gives a total of $2\sum_{i=1}^{n} i = 2\frac{n(n+1)}{2} = n(n+1)$ ways to insert the new leaf edge. □

Proposition A5. At all times in an ERM model, the distribution of ROTs with n leaves is uniform.

Proof. Assume that after k events, all $(m-1)!$ ROTs of size m are equally likely. If the next event is a bifurcation, then, because the result of each (tree, bifurcation event) pair is distinct, after this event all $m!$ ROTs with $m+1$ leaves are equally likely. Similarly, if the next event is an extinction, then for each of the $(m-1)!$ equally likely trees there are m equally likely choices for which leaf to extinguish, giving $m!$ possibilities in all. By Lemma A4, each ROT with $m-1$ leaves results from $m(m-1)$ of these tree-plus-leaf choices. Thus, each ROT with $m-1$ leaves is equally likely, with probability $m(m-1) \times 1/m! = 1/(m-2)!$.

Because this is true for any such sequence of bifurcations and extinctions it is true at all times. □

Of course, any model giving the uniform distribution on ROTs with n tips gives the uniform distribution on rank assignments conditioned on the oriented tree with n tips. Thus, we have the following corollary.

Corollary A6. Any ERM model gives the uniform distribution on rank assignments (and thus tree shuffles) given an oriented tree.

We have the following limited converse of Proposition A5.

Proposition A7. Pure-birth ERM models are precisely the set of pure-birth ranked-oriented-tree processes which, for any $n \geq 1$, give the uniform distribution on ROTs with n taxa when halted as soon as n taxa are present.

Proof. By the proof of Proposition A5, pure-birth ERM models result in a uniform distribution on ROTs of size n (because there have been exactly $n-1$ events).

Now consider a model that does not satisfy the ERM condition. Assume that the kth bifurcation event was not picked uniformly among lineages, that is, there is a ranked tree T_0 with lineages l_1 and l_2 which have probabilities $p_1 \neq p_2$ to speciate. Let T_1 (respectively T_2) be the ranked tree produced if l_1 (respectively l_2) bifurcates. In a pure-birth process, T_1 and T_2 may only be reached in this way,

$$\mathbb{P}[T_1] = \mathbb{P}[T_0] \cdot p_1 \neq \mathbb{P}[T_0] \cdot p_2 = \mathbb{P}[T_2],$$

which shows that this model cannot give the uniform distribution on ranked trees when the process is halted at k taxa. There is only one way to build each ROT with n leaves so the distribution on these cannot be uniform because an equal number must descend from each of T_1 and T_2. Thus, by contradiction, there is no such k and so no such model. □

Note that in the last proposition, the restriction to a pure-birth process is needed. Consider a process with extinction where bifurcation is equally likely for each species, but extinction is history dependent: whenever an extinction event occurs, it undoes the most recent bifurcation event. This model clearly does not belong to the class of ERM models. However, it gives a uniform distribution on ranked trees of some fixed size.

The CRP Model

Proposition A8. A CRP model, stopped at a time depending only on the time and number of leaves, gives the uniform distribution on rank functions for each oriented tree.

Proof. Consider the distribution of ROTs resulting from the stopped CRP. Consider a particular oriented tree, t, with k internal vertices $v_1, \ldots, v_k$. Let n_i and m_i denote the number of internal vertices below the left and right subtrees, respectively, of vertex v_i. Fix a ranking on this tree. We now compute the probability of this ROT under the model (conditional on the total number of leaves). Fix an assignment of p_{v_i} to each internal vertex v_i. Given this choice, the probability of the given ranked tree is the product of the probabilities of each bifurcation event. For a bifurcation at vertex v_i, the probability of this event is the product of p_{v_j} for all v_j for which v_i lies on its left subtree times the product of $(1 - p_{v_j})$ for all v_j for which v_i lies on its right subtree. In the product of these probabilities over all v_i, the term p_{v_j} occurs exactly n_j times (once for each internal vertex on the left subtree of v_j) and the term $(1-p_{v_j})$ occurs exactly m_j times (once for each internal vertex on the right subtree of v_j). Thus, the probability of this ranked tree (given the choice of p_v) is

$$\prod_{j=1}^{k} p_{v_j}^{n_j} \left(1 - p_{v_j}\right)^{m_j}.$$

Note that this is independent of the ranking. Because the p_{v_i} are picked independently from a distribution P, the probability of this ranked tree shape is

$$\int_{p_{v_1}} \cdots \int_{p_{v_k}} \prod_{j=1}^{k} p_{v_j}^{n_j} \left(1 - p_{v_j}\right)^{m_j} \, \mathrm{d}P \cdots \mathrm{d}P,$$

which is again independent of the ranking. Therefore, all rankings of this oriented tree are equally likely. □

Time Complexity of Calculating the Runs Distribution

Here we provide a proof of the time complexity bound for the computation of the runs distribution $\mathcal{R}(T)$ (i.e., conditioning on a given tree shape). This distribution may be computed easily for certain tree shapes, such as the comb tree. However, here we provide a bound that holds for all tree shapes. This bound makes use of a bound on the number of runs in a ranked tree.

Let $r(n)$ denote the maximum number of runs for a ranked tree with n leaves. Thus, $r(1) = r(2) = 0$, $r(3) = 1$, and $r(4) = 2$. Let $I_{i=n/2}$ be 1 if $i = n/2$ and 0 otherwise. For a tree with at least 2 leaves, if the first branch point has i leaves on one side and $n - i$ leaves on the other, with $i \leq n - i$, then the number of runs at this vertex may be up to $2(i-1)+1-I_{i=n/2}$ (note that we have an $(i-1, n-i-1)$ shuffle at this vertex). This maximum is obtained by a shuffle that interleaves the elements from each set, one from each side for as long as possible, starting with the largest side.

Thus, $r(n)$ satisfies the following recurrence: $r(1) = r(2) = 0$ and for $n \geq 2$,

$$r(n) = \max_{1 \leq i \leq n/2} (2i - 1 - I_{i=n/2} + r(i) + r(n-i)).$$

Proposition A9. For all integers $n \geq 1$, $r(n) \leq n \log_2 n$.

Proof. The statement is true for $n = 1$. Suppose that the statement is true for all $k < n$. Then,

$$\begin{aligned} r(n) &= \max_{1 \leq i \leq n/2} (2i - 1 - I_{i=n/2} + r(i) + r(n-i)) \\ &\leq \max_{1 \leq i \leq n/2} (2i - 1 + i \log_2 i + (n-i) \log_2 (n-i)). \end{aligned}$$

Note that $2i - 1$, $i \log_2 i$, and $(n-i)\log_2(n-i)$ are all convex functions of i so their sum is convex also. Thus, the maximum of $2i-1+i\log_2 i+(n-i)\log_2(n-i)$ occurs at an extreme value. Setting $i=1$ gives $1+0+(n-1)\log_2(n-1)$, whereas setting $i=\frac{n}{2}$ gives $2\frac{n}{2}-1+2\frac{n}{2}\log_2\frac{n}{2}=n\left(\log_2 2+\log_2\frac{n}{2}\right)-1=n\log_2 n-1$. Both of these values are less than $n \log_2 n$ and so $r(n)$ must be at most $n \log_2 n$. The result follows for all $n \geq 1$ by induction. □

We now proceed to bound the complexity of computing the distribution of runs for a tree. For a tree T with 1 or 2 leaves, the number of runs is always 0.

Let T be a tree with $n \geq 3$ leaves; we assume a uniform distribution on tree shuffles. Let L and R be the 2 randomly ranked subtrees of T, with a and b leaves, respectively.

Equation (3) may be rewritten as follows:

$$\begin{aligned} &\mathbb{P}\{\mathcal{R}(T) = k\} \\ &\quad = \sum_{i=0}^{A_1} \mathbb{P}\{X_{a,b} = i\} \sum_{j=0}^{A_2} \mathbb{P}\{\mathcal{R}(L) = j\} \mathbb{P}\{\mathcal{R}(R) = k - i - j\} \\ &\quad = \sum_{i=1}^{A_1} \mathbb{P}\{X_{a,b} = i\} \mathbb{P}\{\mathcal{R}(L) + \mathcal{R}(R) = k - i\}, \end{aligned} \tag{A.1}$$

where $A_1 = \min(k, n)$ and $A_2 = \min(k - i, r(a))$. Note that $a + b = n \geq 3$ implies $X_{a,b} \geq 1$ and $\mathcal{R}(T) \geq 1$.

Because $\mathcal{R}(T)$ is supported on (i.e., 0 outside of) $k = 1, \ldots, \lfloor n \log_2 n \rfloor$, and for each k, the computation of equation (A.1) costs $2n - 1$ operations, the cost of computing its distribution with this formula is at most $(\lfloor n \log_2 n \rfloor)(2n - 1)$ arithmetic operations plus the cost of computing $\mathbb{P}\{X_{a,b} = i\}$ for $i = 1, \ldots, n$ and $\mathbb{P}\{\mathcal{R}(L) + \mathcal{R}(R) = x\}$ for $x = 0, \ldots, r(n) - 1 \leq n \log_2 n - 1$.

For these fixed a and b, the values of $\mathbb{P}\{X_{a,b} = i\}$ can be calculated using equation (1) in constant time (at most $5 \times 2 + 4 = 14$ arithmetic operations each) with a linear overhead as follows. The binomial coefficients $\binom{a}{k}$ for $a \leq b$ and $k \leq b$ in equation (1) may be calculated with at most 2 arithmetic operations from the factorials, $j!$ for $1 \leq j \leq n$, which may in turn be precalculated in linear time ($n-1$ multiplications). Thus, calculating $\mathbb{P}\{X_{a,b} = i\}$ for $i = 1, \ldots, n$ takes at most $14n$ arithmetic operations, with a one-time overhead of $n - 1$.

The distribution of $\mathbb{P}\{\mathcal{R}(L)+\mathcal{R}(R)=x\}$ is supported on $x = 0, \ldots, \lfloor n \log_2 n \rfloor - 1$. It may be computed by repeated

application of the formula

$$\mathbb{P}\{\mathcal{R}(L)+\mathcal{R}(R)=x\} = \sum_{j=0}^{\lfloor (n-1)\log_2(n-1)\rfloor} \mathbb{P}\{\mathcal{R}(L)=j\}\mathbb{P}\{\mathcal{R}(R)=x-j\}$$

as long as the distributions of $\mathcal{R}(L)$ and $\mathcal{R}(R)$ are known. This computation requires at most $n\log_2 n(2(n-1)\log_2(n-1)+1)$ arithmetic operations, and at most $(n-1)\log_2(n-1)+1$ multiplications and $(n-1)\log_2(n-1)$ additions for each of $n\log_2 n$ values of x. Note that the distribution of $\mathbb{P}\{\mathcal{R}(L)\}$ is supported by $j = 0, \ldots, \lfloor (n-1)\log_2(n-1)\rfloor$ because L has at most $n-1$ leaves.

So, if the distribution of $\mathcal{R}(L)$ and $\mathcal{R}(R)$ are known, the distribution of $\mathcal{R}(T)$ may be calculated in at most

$$(\lfloor n\log_2 n\rfloor)(2n-1)+14n+n\log_2 n\,(2(n-1)\log_2(n-1)+1)$$

arithmetic operations. This is at most

$$2n^2\log_2 n + 2n^2\log_2^2 n + 14n$$

for all $n \geq 3$. Because $\mathcal{R}(T)$ is 0 for $n = 1, 2$, the time to calculate it is 0.

This procedure may be applied recursively, computing the distribution of runs of all subtrees before finally computing the runs distribution of T. Because there are $n-1$ internal vertices and each has at most n leaves below it, the total number of arithmetic operations required is at most $n(2n^2\log_2 n + 2n^2\log_2^2 n + 14n + 1)$ (including the overhead for precomputing j!). This is $O(n^3\log_2^2 n)$.